TILBURG UNIVERSITY

Tilburg University

Multi-source statistics

de Waal, Ton; van Delden, Arnout; Scholtus, Sander

Published in:
International Statistical Review

DOI:
10.1111/insr.12352

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
de Waal, T., van Delden, A., & Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review*, *88*(1), 203-228. https://doi.org/10.1111/insr.12352

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

International Statistical Review

International Statistical Review (2019), 0, 0, 1–26 doi:10.1111/insr.12352

Multi-source Statistics: Basic Situations and Methods

Ton de Waal[1,2], Arnout van Delden[1] and Sander Scholtus[1]

[1]*Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands*
[2]*Department of Methods and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands*
E-mail: t.dewaal@cbs.nl

Summary

Many National Statistical Institutes (NSIs), especially in Europe, are moving from single-source statistics to multi-source statistics. By combining data sources, NSIs can produce more detailed and more timely statistics and respond more quickly to events in society. By combining survey data with already available administrative data and Big Data, NSIs can save data collection and processing costs and reduce the burden on respondents. However, multi-source statistics come with new problems that need to be overcome before the resulting output quality is sufficiently high and before those statistics can be produced efficiently. What complicates the production of multi-source statistics is that they come in many different varieties as data sets can be combined in many different ways. Given the rapidly increasing importance of producing multi-source statistics in Official Statistics, there has been considerable research activity in this area over the last few years, and some frameworks have been developed for multi-source statistics. Useful as these frameworks are, they generally do not give guidelines to which method could be applied in a certain situation arising in practice. In this paper, we aim to fill that gap, structure the world of multi-source statistics and its problems and provide some guidance to suitable methods for these problems.

Key words: administrative data; data integration; multi-source statistics; statistical methods; survey data.

1 Introduction

Many National Statistical Institutes (NSIs), especially in Europe, are moving from single-source statistics to multi-source statistics. This is due to higher quality demands with respect to the statistics produced: more detailed data, more timely data and a general demand for a faster response from NSIs to events in society. In addition, many NSIs face budget cuts that make large-scale surveys too costly to set up and maintain.

National Statistical Institutes traditionally have produced single-source statistics, where basically only data from a single data source are utilised. Other data sources are often used in this process too, but only as auxiliary data, for instance, to calibrate or improve estimates, or as supplemental data to validate the statistics produced. In most cases, the single data sources are surveys, although nowadays administrative data are more and more used as single data sources and also Big Data are starting to be used (see, e.g. Daas *et al.*, 2015; Landefeld, 2014).

By combining data sets, more detailed statistics can be produced. By utilising a combination of already available data sets, NSIs can also produce more timely statistics and respond more quickly to events in society, as one does not have to wait until these data have been collected.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute. Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

By combining survey data with already available administrative data and Big Data, NSIs can reduce data collection and processing costs and reduce the burden on respondents.

Moving from single-source to multi-source statistics therefore seems the way to go. However, this transition is not an easy one. Multi-source statistics come with new problems that need to be overcome before the resulting output quality is sufficiently high and before those statistics can be produced efficiently. What complicates the production of multi-source statistics is that supporting data come in many different varieties as data sets can be combined in many different ways. Every variety seems to come with its own problems for which tailor-made solutions are needed. It often feels like for every new multi-source statistics one has to reinvent the wheel.

Given the rapidly increasing importance of producing multi-source statistics in Official Statistics, there has been considerable research activity in this area over the last few years. Some frameworks have been developed for multi-source statistics; see, for instance, Bakker and Daas (2012) and Zhang (2012), who focus on processing steps and error sources in multi-source statistics. Useful as these frameworks are, they generally do not give guidelines to which method could be applied in a certain situation arising in practice.

In the current paper, we do not strive to offer an all-encompassing theoretical framework of some kind, such as a framework attempting to describe all possible situations. Instead, this paper has a more pragmatic aim. Our goal is to provide practical guidelines for producers of multi-source statistics on which issues may be encountered and which kinds of methods can be applied to overcome these issues in practice. In order to identify the most important research questions with respect to multi-source statistics, we propose a breakdown into eight basic situations that seem to be most commonly encountered in practice.

The remainder of the paper is organised as follows. Section 2 discusses some characteristics of multi-source statistics. These characteristics can be used to identify basic situations for multi-source statistics. Section 3 focuses on some general issues when combining multiple data sets. Section 4 describes eight important basic situations in detail, as well as corresponding methodological challenges and methods to overcome these challenges. Section 5 concludes the paper with a discussion.

2 Characteristics of Situations for Combining Data

The characterisation of situations for combining multiple data sets can be complicated due to the inherent heterogeneous nature of the data. For these situations, both input and output characteristics are of importance. The input characteristics determine the data availability whereas the output characteristics set the target for which the data are combined. The latter is important for deciding which methods can be used. We first discuss the input and then the output characteristics.

2.1 Characteristics of the Inputs

Within the input characteristics, there are three fundamental dimensions: the representation dimension, the measurement dimension and the time dimension (Nordbotten, 2010). In addition to this, also the aggregation level of the data is relevant. This is less fundamental, because it is a relative term. For instance, stratum totals of economic statistics are aggregates of the underlying micro data, but these totals can be considered as low aggregate values when they enter National Accounts. Likewise, output from NSIs can be considered low aggregate values in European Statistics.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

For each dimension and for the aggregation level, one or more aspects are important, given in the tables below. Each aspect can have multiple 'states'; for instance, the aspect 'population' can have two states: we know the population—for instance, because that information is available from a population register (frame) or from a Census—or we do not know the population.

We present the representation dimension in Table 1, the measurement dimension in Table 2 and the time dimension in Table 3.

Finally, for the aggregation level, we distinguish three different states (Table 4).

Table 1. *Representation dimension.*

Population	Unit selections		Coverage	Unit distinctness
	Individual data sets	Combined data	(with respect to target population)	
1. The set of population units is known 2. The set of population units is not known	1. The data set contains a complete enumeration of its target population 2. The data set is selected by means of probability sampling from its target population 3. The data set is selected by non-probability sampling from its target population	1. Together the data sets contain a complete enumeration of the target population 2. Together the data sets do not contain a complete enumeration of the target population	1. The data contain no undercoverage and no overcoverage 2. The data contain undercoverage but no overcoverage 3. The data contain overcoverage but no undercoverage 4. The data contain both undercoverage and overcoverage	1. There are no overlapping units in the data sets 2. (Some of the) units in the data sets overlap

Table 2. *Measurement dimension.*

Completeness	Variable distinctness	Relatedness
1. Together the data sets contain all target variables 2. Part of the target variables need to be derived from the source variables	1. There are no overlapping variables in the data sets 2. There are no overlapping target variables in the data sets, but there are overlapping auxiliary variables 3. (Some of the) target variables in the data sets overlap, but the concepts are measured in different ways 4. (Some of the) target variables in the data sets overlap, and the concepts are measured in the same way	1. There are no logical relations between variables in different data sets 2. There are logical relations between variables in different data sets (hard or soft constraints)

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Table 3. *Time dimension.*

Repeated measures	Time reference	Availability	Progressiveness
1. The data are cross-sectional (time stamp or period) 2. The data are longitudinal	1. The data refer to a single time point or period 2. The data refer to events (transitions between periods)	1. The data set contains all data for all units from first availability 2. The data set does not contain all data for all units from first availability but becomes gradually available over time	1. The data values in the data set are final 2. The data values in the data set are updated over time

Table 4. *Aggregation level.*

1. The data sets consist of only micro data.
2. The data sets consist of a mix of micro data and aggregated data.
3. The data sets consist of only aggregated data

Table 5. *Characteristics of targeted output.*

Type of output	Usage of data sets	Quality improvement of processing
1. The output concerns micro data sets 2. The output concerns population registers 3. The output concerns statistics 4. The output concerns metadata	1. Estimates are obtained by direct tabulation from micro data 2. Estimates are indirectly obtained by more complex estimation methods	1. Achieve relevant estimates 2. Achieve accurate and reliable estimates 3. Achieve timely and punctual estimates 4. Achieve coherent and comparable estimates 5. Achieve accessible and clear estimates

2.2 Characteristics of Targeted Output

We now turn towards the output characteristics. For the targeted output, three different aspects are important: the type of output, the usage of data sets and the main quality improvement that is intended by data processing. For each of those aspects, different states are relevant, given in Table 5. The states of the aspect 'quality improvement' refer to the five quality dimensions that are distinguished in Eurostat (2015, pp. 21–107).

In the present paper, we limit ourselves to descriptive statistics, such as totals and means, as output. In particular, we will assume that the main aim of multi-source statistics is to produce high-quality estimates at an aggregated level.

The total number of possible states when combining only two data sets is already very large. To give a first idea, assume that both data sets concern micro data on events. We would then obtain the following combinations of states: 'population' (2 states) ×'unit selections' (3 states) ×'coverage' (4 states) ×'unit distinctness' (2 states) ×'completeness' (2 states) ×'variable distinctness' (4 states) ×'relatedness' (2 states) ×'repeated measures' (2 states) ×'availability' (2 states) ×'progressiveness' (2 states) = 6 144 potential states. We have omitted multiplication

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Table 6. *Overview of eight basic situations (integration problems) of multi-source statistics and their characteristics.*

No.	Integration problem	Characteristics of targeted output		Characteristics of input sources			
		Usage of sources	Quality improvement of processing	Representation dimension	Measurement dimension	Time dimension	Aggregation level
1	How to deal with complementary variables?	Direct tabulation	Relevant estimates	Complete enumeration, Overlapping units	Overlapping auxiliary variables	Cross-sectional	Only micro data
2	How to deal with complementary units?	Direct tabulation	Relevant estimates	Complete enumeration, No overlapping units	Overlapping target variables	Cross-sectional	Only micro data
3	How to estimate the joint distribution of variables in non-overlapping samples?	Direct tabulation	Relevant estimates	Not a complete enumeration, No overlapping units	Overlapping auxiliary variables	Cross-sectional	Only micro data
4	How to estimate true values of variables with conflicting values?	Indirect tabulation	Accurate and reliable estimates	Overlapping units	Overlapping target variables	Cross-sectional	Only micro data
5	How to estimate population size?	Indirect tabulation	Accurate and reliable estimates	Not a complete enumeration, Overlapping units	Overlapping auxiliary variables and/or overlapping target variables	Cross-sectional	Only micro data

Continues

International Statistical Review (2019), **0**, 0, 1–26
© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Table 6. *Continued*

No.	Integration problem	Characteristics of targeted output		Characteristics of input sources			
		Usage of sources	Quality improvement of processing	Representation dimension	Measurement dimension	Time dimension	Aggregation level
6	How to make new estimates consistent with previously published estimates?	Indirect tabulation	Coherent and comparable estimates	*	Overlapping target variables	Cross-sectional	Mix of micro data and aggregated data
7	How to achieve numerical consistency in accounting equations?	Indirect tabulation	Coherent and comparable estimates	*	*	Cross-sectional	Only aggregated data
8	How to achieve numerical consistency between low and high frequency data?	Indirect tabulation	Coherent and comparable estimates	*	Overlapping target variables	Longitudinal	Only aggregated data

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

with the two states of 'combined unit selections' because that partly follows from the unit selections in the individual data sets. We also omitted 'time reference', because event data are often only longitudinal.

It is clear that we cannot describe all possible different situations. In the remainder of this paper, we have limited ourselves to eight often occurring 'basic' situations in combining data sets in official statistics. Besides being situations that often occur in practice, each of them also illustrates certain problems that can arise when combining data sets. That these eight situations indeed cover most situations occurring in official statistics is confirmed by feedback we received on presentations at various conferences (e.g. NTTS conference 2017; see De Waal, Van Delden and Scholtus, 2017b) and workshops.

Table 6 provides an overview of the eight basic situations together with 'defining states' for each of these basic situations. An asterisk (*) in Table 6 denotes that for that basic situation, the characteristic is not a 'defining state'.

3 General Issues when Combining Data

Two issues apply to many situations where data sets are combined: harmonisation and record linkage. Both units and variables in the various data sets may need to be harmonised before these data sets can be combined. An important reason for harmonisation is the so-called unit error problem. Unit errors occur when units are defined differently in one data set than in another data set, when the units in available data sets are not defined according to the official definition that one wants to use at the NSI or when units have to be constructed. In the Netherlands, for instance, administrative units for value added tax (VAT) data may differ from administrative units for profit and loss data. In turn, those administrative units may differ from the statistical units for which the target population is defined. A specific version of the unit problem occurs when data are available at different levels of aggregation only. For instance, we may want to combine data on bankruptcies (available at the level of legal persons) with data on the number of jobs of employees. The latter are available at the level of enterprises, where an enterprise may be a combination of legal persons. For more details on the unit error problem, we refer to Zhang (2011, 2012) and Van Delden *et al*. (2018a).

Target variables in the data sets may also need to be harmonised. For example, in the Netherlands, quarterly turnover of enterprises available from administrative data obtained from the tax office often differs from quarterly turnover available from a survey. An important special case that requires harmonisation of variables occurs when we have a subset of the variables in one data set (say administrative data) and other variables in a second data set (say sample survey data) and the sets contain overlapping units, but the reference periods of the two sets are different. For many variables, values differ for different reference periods.

A closely related harmonisation issue is timeliness of data sets. Different data sets may be available at different moments, and the quality of the data sets may vary over time. In particular, the progressiveness of administrative data, that is, the fact that administrative data sets generally contain more and/or higher quality data as time passes, often presents a problem for early estimates (see also Zhang, 2014). The problem that part of the data are at first missing may in some cases be solved by means of weighting (see, e.g. Särndal *et al*., 1992) or imputation techniques (see, e.g. Little and Rubin, 2002). A complicating aspect is that the initially observed data may be from a selective part of the population (see, e.g. Ouwehand and Schouten, 2014, for assessing the representativeness of data). A further issue is that corrections on originally reported data may become available after a long time. Measurement error in earlier versions of

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

the data may in some cases be treated by measurement error correction methods, for instance, methods as discussed in Section 4.4.

Micro-integration is often a first step to harmonise units and variables (see, e.g. Bakker, 2011a, 2011b). In micro-integration, for instance, rules may be used to derive the target variables from those present in the input data sets. Micro-integration cannot solve all the harmonisation problems that arise in the context of multi-source statistics, and more advanced methods are often required; see Sections 4.2 and 4.4.

The second common issue is record linkage. We need a record linkage step to link the units in the data sets to the population register or to each other. When unique unit identifiers, such as unique personal identification numbers, are present in the data sets, deterministic linkage can be used (see, e.g. Chapter 8 in Herzog *et al*., 2007). When the same non-unique identifier variables, such as names and addresses, are present in both sets, probabilistic linkage (see, e.g. Fellegi and Sunter, 1969) or machine-learning based record linkage (see, e.g. Christen, 2012) might be used.

Misspelling and variation of formats of, for instance, names and addresses, can severely complicate the record linkage process. As a result, correct matches may be missed in the record linkage process ('false negatives'), and incorrect matches may be made ('false positives'). Such 'false negatives' and 'false positives' may lead to bias in estimates based on linked data and may hamper the analysis of linked data. Some methods have been proposed that aim to correct these biases for record linkage error. For more details on the issues of record linkage, the effects of record linkage error on estimates and the analysis of linked data, and on methods to correct for record linkage error, we refer to Harron *et al*. (2016), especially Chapters 1, 4, 5 and 6.

Record linkage becomes even more problematic in the case of unit errors, which emphasises the important role of harmonisation.

4 Basic Situations and Their Methods

In this section, we present eight basic situations that we consider to be the most important ones in practice (see also Table 6). We propose and elaborate these basic situations with respect to the aspects mentioned in Section 2. Many practical situations can be built on these basic situations.

We use figures to illustrate the eight basic situations. Concerning these illustrations, we note that the white rectangle to the left represents the population frame with units; the two light grey colours ( , ) represent different input data sets and the dark grey colour () represents derived output statistics; blocks with horizontal line patterns represent aggregated data and blocks without a filling pattern represent micro data. The arrow refers to the complete process to go from input data to output statistics. In some basic situations, specific methodology is needed as part of this process, and in those cases, the methodology is mentioned in the corresponding section. The target variables in the data sets, denoted by $Y_1, \ldots, Y_p$, are observed for units $1, \ldots, N$ in the case of a full enumeration of the population, or observed for units $1, \ldots, n$ with $n < N$ in the case of a sample. The general notation for the corresponding target parameters is $\hat{\theta}_1, \ldots, \hat{\theta}_p$. In practice, these will often be estimated for a set of domains $h = 1, \ldots, H$ within the population. For clarity of presentation, those domains are omitted in most of the figures. Further, background variables, denoted by $\mathbf{Z} = (Z_1, \ldots, Z_k)'$ may play a role in the methodology to link the data sets. Background variables are omitted from the figures unless they are a crucial part in the estimation procedure of the target parameters. In some figures, specific symbols are used, which are explained in the corresponding basic situation.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

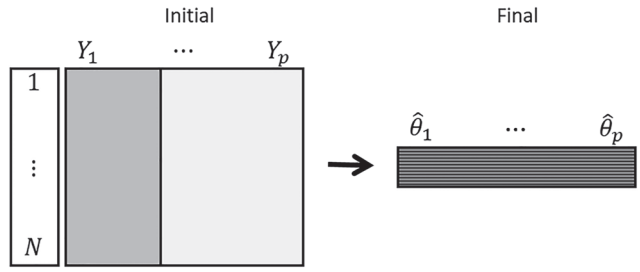

Figure 1. *Combining micro data sets with full population coverage and complementary target variables.*

4.1 Data Sets with Full Population Coverage and Complementary Target Variables

The first basic situation concerns multiple cross-sectional micro data sets covering the target population where the different data sets contain complementary target variables (see Figure 1). We refer to this as the 'split-variable' case. Provided that the data are error-free, the data can simply be linked to produce output statistics.

Figure 1 illustrates the situation that we are interested in: estimating a set of p target parameters based on variables that are observed for all N units of the population or for a probability sample of size $n < N$. The sampling case may be less common for Situation 1 than for the other situations, but it may occur when linking a sample survey to register data.

In this situation, record linkage is an important issue. We assume that the data sets also contain a set of background variables **Z**, for instance, variables that are used to link the data sets to the population register.

An example of Situation 1 is the integration of different administrative data sets on economic performance of businesses. For instance, in the Netherlands, administrative data on profit and loss are sometimes combined with administrative data on personnel costs.

An example of unit type differences and linkage issues occurred in the integration of various administrative data sets at Statistics Netherlands to compute energy use per meter squared for dwellings and for businesses or institutions. The central data concern administrative client energy data sets (CAD) obtained from gas and electricity distributors, which consist of the complete volume of energy delivery in the Netherlands. The CAD is linked to a central register on addresses and buildings (Kadaster), which contains building/dwelling type and their area. It is also linked to a general business register (GBR) to identify business activities and to find the economic activity. The unit type within the CAD is the 'energy connection point', identified by a unique energy connection point number (Dutch: EAN). The EAN is related to an address and client name. This address information is also found in the Kadaster data and in the GBR data.

The linkage by address is not always one-to-one. One address may contain multiple energy connection points, which can be solved by adding up the energy use of the different EANs. In addition, one may also have one EAN that is linked to a building that contains multiple activities/enterprises. In this case, one often appoints the energy use to the dominant activity in that building, which is not an ideal approach. One alternative is to introduce explicit categories expressing the mixture, such as 'building with multiple business activities'. Another alternative is to try to estimate the energy use per economic activity by using auxiliary variables. For instance, a linear regression model could be constructed with size class, economic activity per enterprise and floor surface per enterprise as explanatory variables. An example of a similar approach can be found in Enderer (2008). If the model predictions are reasonably accurate, we prefer the use of a model in this and similar situations.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

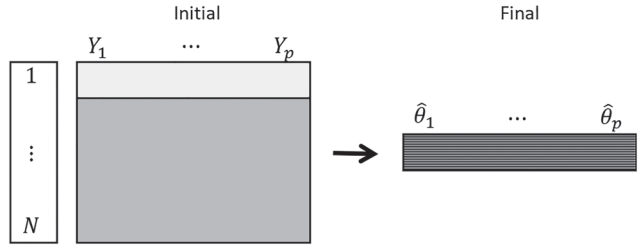

Figure 2. *Combining complementary micro data sets that, together, have full population coverage.*

4.2 Data Sets with Full Variable Coverage and Complementary Units

The second basic situation also concerns multiple cross-sectional micro data sets covering the target population, but in this case, the different data sets contain different units (see Figure 2). We refer to this as the 'split-population' case. Provided the data are in an ideal error-free state and the concepts are identical, the different data sets are *complementary* to each other in this case, and likewise to Situation 1, they can be simply 'added' to each other in order to produce output statistics. However, in practice, often a harmonisation step will be necessary to correct for differences in the conceptual definitions of the variables.

An example of Situation 2 is the estimation of quarterly turnover at Statistics Netherlands. The turnover data are available from a combination of census data and administrative VAT data, and both are linked to the GBR (Van Delden and De Wolf, 2013). The VAT data are available for fiscal administrative units, and they can be uniquely linked to the enterprises in the GBR only for the small and medium sized enterprises. The complementary group of large and complex enterprises receives a census survey. Statistics New Zealand (Chen *et al*., 2016) uses a very similar approach, where sub-annual sales data are obtained from administrative Goods and Service Tax data, complemented by survey data for the large and complex units.

A method to harmonise variables based on multiple data sets that relies on the assumption that one data set can be used as the 'gold standard' is given in Van Delden *et al.* (2016). They analysed the relation between the metadata and the data of annual survey turnover and VAT in 2009 and 2010, where survey turnover was considered to be the 'gold standard'. The relation was analysed for more than 300 domains of economic activity. They divided the domains into four groups. The *Control* group concerned domains where there are no conceptual differences in the definitions of survey and VAT turnover. These domains showed a linear relationship with an intercept close to 0 and a slope that was very close to 1. The *Accept* group concerned domains with conceptual differences but only small numerical differences. The *Adjust* group concerned domains with conceptual differences and systematic numerical differences. For the units in this domain, a correction factor can be applied to estimate the survey turnover values from the VAT turnover values. The final group, *Reject*, concerned domains with conceptual differences and large non-systematic numerical differences. For units in the *Reject* group, VAT data cannot be used, and we have to continue using survey data. For units in the *Control*, *Accept* and *Adjust* groups, the survey can be abolished. Examples of the relations between survey and VAT turnover can be found in Figure 2D–F in Van Delden *et al.* (2016).

Another example of Situation 2 is where national figures are composed from a large set of decentralised, autonomous administrations, for instance, national health care figures based on regional administrations. In such situations, hierarchical models may be of use where the mean of each decentralised system is modelled as a random effect and the individual records are nested within each separate source. The challenge there is to account for a bias component

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

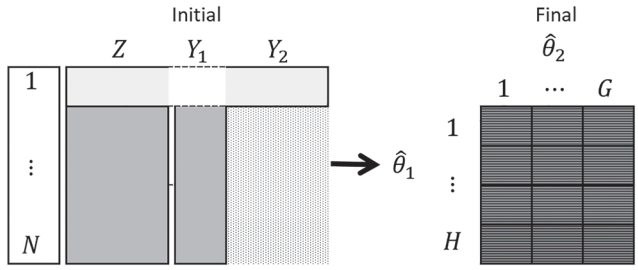

Figure 3. *Combining non-overlapping micro data sets with part of the variables in a single source, with full population coverage. The data sets can be samples from the population.*

per data source (see Lohr and Raghunathan, 2017, and references therein; Van Delden *et al*., 2018b).

4.3 Overlapping Variables but Non-overlapping Units

A slightly different situation occurs when, besides having non-overlapping units as in Situation 2, we also have a number of overlapping variables and some target variables that are available in only one of the data sets. We call this Situation 3 (see Figure 3). We still would like to join the target variables $\mathbf{Y}_1$ in one of the data sets to target variables $\mathbf{Y}_2$ in another data set and estimate the joint distribution of variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (represented by the rectangle in Figure 3, where the estimates of both variables are divided into different classes). For this, statistical matching techniques are available.

In Italy, the main data sets available for estimating household income and expenditure are the Household Budget Survey conducted by the Italian National Institute of Statistics and the Survey on Household Income conducted by the National Bank of Italy. Unfortunately, there is no single data set available that contains data on both household income and expenditure. In order to examine the effects of policy changes on the relation between household income and expenditure, one therefore resorts to using statistical matching (see Conti *et al.*, 2017).

Statistical matching differs fundamentally from record linkage. Whereas in record linkage one aims to link a record from a unit in one data set to a record from the *same* unit in another data set, in statistical matching, one essentially aims to match a record of a unit in one data set to a record from a *similar*, but generally not the same, unit in another data set.

Statistical matching can be carried out at the micro level or at the macro level. When statistical matching is carried out at the micro level, one combines data from individual units in the different data sets to construct synthetic records with information on all variables. In particular, when there are two data sets, information from one data set, the donor data, is used to estimate target values in the other data set, the recipient data. The records constructed are a mix of data from different units from different data sets.

When statistical matching is carried out at the macro level, one assumes a parametric model for all the data, for instance, a multivariate normal model for numerical data, and then estimates the parameters of this model. These parameters are subsequently used to estimate the population parameters one is interested in. For an overview of methods for statistical matching at both the macro level and the micro level, we refer to Chapters 2 and 3 in D'Orazio *et al*. (2006).

In Figure 3, we have two data sets. Data set 1 contains variables $\mathbf{Y}_1$ and $\mathbf{Z}$ and data set 2 $\mathbf{Y}_2$ and again $\mathbf{Z}$. Variables $\mathbf{Z}$ are the common (background) variables that are used to statistically match the records. When statistical matching is carried out at the micro level, variables $\mathbf{Z}$ are used to match individual units in data set 1 to individual units in data set 2.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

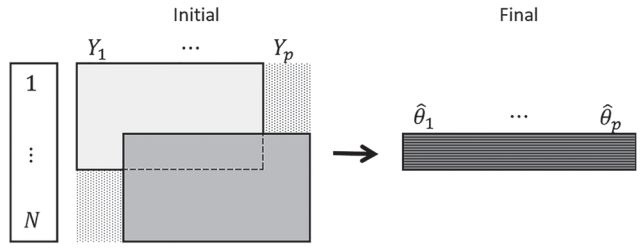

Figure 4. *Combining overlapping micro data sets with full population coverage.*

The fundamental issue of statistical matching is that the relationship between the target variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ cannot be estimated directly, but only indirectly. In order to do so, one has to rely on untestable assumptions, that is, untestable from the data sets themselves, about this relationship. The most common assumption is the conditional independence assumption (CIA), which says that conditional on the values of background variables $\mathbf{Z}$, the target variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent. In general, the joint relationship between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ can be decomposed into a part which is explained by $\mathbf{Z}$ and a remaining part which is unexplained by $\mathbf{Z}$. In the simple case of a trivariate normal distribution, this can be written as $\sigma_{Y_1Y_2} = \sigma_{Y_1Z}\sigma_{Y_2Z}\big/\sigma_Z^2 + \sigma_{Y_1Y_2|Z}$ (see Stuart and Ord, 1991, pp. 1010–1011). If the CIA holds, then $\sigma_{Y_1Y_2|Z} = 0$.

As an alternative to the CIA, the so-called instrumental variable assumption has recently been proposed (see Kim *et al.*, 2016). An instrumental variable is a variable that induces changes in the target variable of one data set but has no effect on the target variable of the other data set. In practice, it may be hard to find such a variable.

When the total output uncertainty based on the CIA or instrumental variable assumption is too large, one can make use of auxiliary data (Singh *et al.*, 1993). One option is to link an administrative variable to both data sets. Van Delden *et al*. (2019) found that even when the administrative variable is strongly related to a target variable in one of the data sets, the resulting uncertainty is often too large to be useful in official statistics. Alternatively, one might use a third data set where the common variables and the target variables in the two data sets are observed. This third data set can be obtained from a population that is close to the target population (a proxy) or it can concern data from a small overlap of the two data sets. The use of such a third data set would lead to Situation 4, which is discussed in the next section.

4.4 Overlapping Variables and Overlapping Units

Situation 4 (see Figure 4) is characterised by a deviation from Situation 2, by which there exists an *overlap* concerning both units and measurements between the different data sets.

In this situation, at least for a subset of the units in the population, we have multiple measurements of the same target variable(s), coming from different data sets. Due to measurement and timing errors, these observed variables from different sets will usually not agree exactly for all units. An example of Situation 4 arises in education statistics in the Netherlands. There exist both administrative and survey data on the education level of Dutch people (Linder *et al*., 2011). Some persons can be found in both data sets, and the respective education level measurements do not always agree with each other as both sets may contain measurement errors.

When the same phenomenon is observed for the same units in multiple data sets, one can utilise the multiple observations to identify and correct residual errors. An approach that is often used in practice at NSIs is micro-integration (Bakker, 2011a). In addition to the harmonisation step described in Section 3, in the present situation, micro-integration also involves comparing

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

the available observations for each overlapping unit to determine which of the data sets is most likely to contain the best approximation of the true value for that unit. Often, deterministic correction and derivation rules are used for this. In many applications, some form of micro-editing is also needed to obtain consistency between different target variables observed in different data sets (Di Zio and Luzi, 2014; De Waal *et al*. 2011).

Micro-integration is a rather crude and somewhat subjective technique. It can be used to harmonise the most important and most obvious inconsistencies between data sets, but not to harmonise more subtle inconsistencies. When such more subtle inconsistencies are caused by measurement error, it may in some cases be possible to find an appropriate statistical model for the measurement errors in the observed variables. Model-based estimates can then be obtained for the underlying true values of the target variable(s), either at the individual level or directly at the level of the target parameters. The true value itself is (usually) not observed; this is called a latent variable. The precise relation between the latent true value and the observed values depends on the type of model. In their basic form, most measurement error models assume that the errors are independent across observed variables, given the underlying true value; this is known as the local (or conditional) independence assumption.

To model measurement errors in numerical data, one may use a structural equation model (e.g. Bollen, 1989) or a finite mixture model (e.g. McLachlan and Peel, 2000). Recently, applications of structural equation modelling to multi-source statistics have been considered by Bakker (2012) and Scholtus *et al*. (2015). Finite mixture models have been developed by Meijer *et al*. (2012) and Guarnera and Varriale (2015, 2016). Under such a model, the population is supposed to consist of two or more components where each component has a different distribution of observed values, and each unit is supposed to belong to one of these components. Guarnera and Varriale explicitly consider the case that measurement errors are 'intermittent': part of the observed values in each data set are correct, and the remaining values contain errors.

For categorical data, models based on latent class (LC) analysis can be used (e.g. Hagenaars and McCutcheon, 2002). Application of LC models to measurement errors in statistical data are considered by, among others, Biemer (2011), Si and Reiter (2013), Pavlopoulos and Vermunt (2015) and Oberski (2017).

Boeschoten *et al*. (2017) also use an LC model to model the true value of a variable that is observed (with measurement error) in multiple sources. We sketch their approach. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_s)'$ denote a vector of observed categorical variables that measure the same conceptual variable of interest (e.g. in s different data sources). The true value with respect to the variable of interest is represented by a latent variable X. We assume that all variables Y_j and X have the same set of categories, say $1, \ldots, L$. Under the local independence assumption, the marginal probability $\Pr(\mathbf{Y} = \mathbf{y})$ of observing the particular vector of values $\mathbf{y} = (y_1, y_2, \ldots, y_s)'$ can be expressed as

$$\Pr(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{L} \Pr(X = x) \prod_{j=1}^{s} \Pr\big(Y_j = y_j \mid X = x\big).$$

Estimating the LC model amounts to estimating the probabilities in the right-hand-side of this expression. The model can be used to estimate, for each unit in the data, the probability of belonging to a particular LC, given its vector of observed values:

$$\Pr(X = x \mid \mathbf{Y} = \mathbf{y}) = \frac{\Pr(X = x) \prod_{j=1}^{s} \Pr(Y_j = y_j | X = x)}{\sum_{x'=1}^{L} \Pr(X = x') \prod_{j=1}^{s} \Pr(Y_j = y_j | X = x')}. \tag{1}$$

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

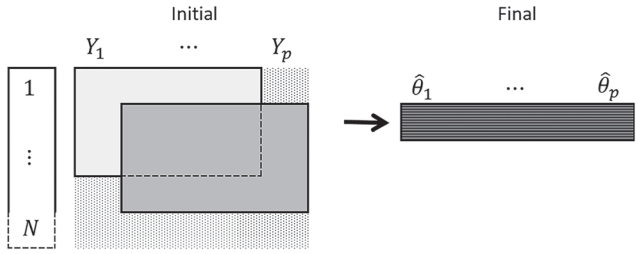

Figure 5. *Combining overlapping micro data sets with undercoverage.*

The method proposed by Boeschoten *et al*. (2017) starts with the original combined data set and then proceeds with five steps.

1. Select m bootstrap samples from the original combined data set.
2. Create an LC model for every bootstrap sample.
3. Multiply impute latent 'true' variable X for each bootstrap sample. m empty variables $(W_1, \ldots, W_m)$ are created and imputed by drawing one of the categories using the estimated posterior membership probabilities (1) from the m LC models.
4. Obtain estimates of interest from the imputed variables.
5. Pool the estimates using Rubin's rules for pooling (see Chapter 3 in Rubin, 1987, p. 76). An essential aspect of these pooling rules is that an estimated variance of the pooled estimates is obtained.

The method is, besides the local independence assumption, based on two additional assumptions: that measurement errors are independent of the covariates and that covariates do not contain classification errors. When covariates do contain classification error, the method can lead to biased estimates.

Estimated relations between the target variable and covariates are only valid when these covariates are taken into account in the LC model and if there is not too much measurement error in the underlying data sets. If covariates are not taken into account in the LC model, either a new LC model needs to be estimated and applied or a correction method should be used (see, e.g. Boeschoten *et al*., 2018).

A related method for correcting for measurement error is multiple over-imputation, where data affected by measurement error are multiply imputed (see, e.g. Blackwell *et al*., 2017). In contrast to imputation, with over-imputation observed values may be replaced by imputed values. Van der Heijden *et al.* (2018) proposed an imputation approach for the case where the measurements of a target variable in one data set are considered to be of higher quality than the measurements of that variable in other data sets, and some values in the higher quality data set are missing.

Before applying a structural equation model, LC model or imputation model, large errors in the data usually need to be corrected by micro-integration or a form of micro-editing.

4.5 Undercoverage and Overcoverage

Situation 5 is characterised by a further deviation from Situation 4, by which the combined data entail *undercoverage* of the target population, even when the data are otherwise in an ideal error-free state (see Figure 5). In this situation, the total population size is not known.

Producers of official statistics are often interested in estimating the unknown size of a population. In particular, an important problem in a population census is to estimate the number of

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

persons in the target population who were missed by all data sets used in the census. The so-called capture–recapture methods are often used to solve this problem (Fienberg, 1972; Chapter 6 in Bishop *et al*. 1975; International Working Group fror Disease Monitoring and Forecasting, 1995).

The simplest application of the capture–recapture method is based on two independent samples from the target population. Consider a 2×2 contingency table with the observed counts of persons being included or excluded in the first and second sample. Let n_{11} denote the observed number of persons in the overlap of the two samples, and let n_{10} and n_{01} denote the numbers of persons observed in the first sample but not the second sample and vice versa. By definition, one does not observe any persons that are not in either sample ($n_{00} = 0$). Let m_{00} denote the expected number of persons in the population that are not observed in either sample. If the samples are independent, a consistent estimator for m_{00} can be obtained from the observed counts as follows (e.g. Bishop *et al.*, 1975, p. 232): $\hat{m}_{00} = n_{10}n_{01}/n_{11}$. An estimate for the total population size, including the part that was missed by both samples, is then given by $\hat{N} = n_{11} + n_{10} + n_{01} + \hat{m}_{00}$. Formally, the capture–recapture method can be derived from a log-linear model for the aforementioned contingency table (see Chapter 6 in Bishop *et al.*, 1975). This approach is also referred to as dual system estimation (Ding & Fienberg, 1994)

An example of Situation 5 where the capture–recapture method can be applied concerns a population census followed by a post-enumeration survey (Wolter, 1986; Brown *et al.*, 1999; Brown *et al*. 2006). Here, the post-enumeration survey is conducted with the specific aim of estimating the undercount in the original population census. The capture–recapture method can also be applied by NSIs that conduct a census based on administrative data (Van der Heijden *et al.*, 2012; Baffour *et al*. 2013; Gerritse, 2016). In this case, data from at least two administrative sources are linked together, and each data set is considered as an independent sample from the population.

Gerritse *et al.* (2016) applied a capture–recapture method to estimate the amount of undercoverage in the population size estimate of the 2011 Dutch census, which is a virtual census in the sense that it is mainly based on a number of administrative data sets, supplemented with sample survey data. The census itself was based on the Dutch population register. For the estimation of undercoverage, two additional registers were linked to the population register: an employment register and a crime suspects register. The census aims to count the number of 'usual residents', where persons are classified as usual residents if they have lived at least 12 months in the Netherlands or intend to do so at the time of the census. Gerritse *et al.* (2016) used probabilistic linkage to link the three registers. To handle missing values on the 'usual resident' status, two different approaches were used: maximum likelihood estimation and imputation by predictive mean matching. The latter approach was found to be more flexible and therefore preferred by the authors.

The capture–recapture method is based on five assumptions (Gerritse, 2016):

(a) The event of being observed in one data set should be independent of the event of being observed in the other data set. This assumption can be relaxed if there are three or more sources (see Chapter 6 in Bishop *et al*., 1975) or by adding covariates to the model (Van der Heijden *et al*., 2012; 2018).
(b) The target population should not change during the period of observation in each data set (i.e. the population should be 'closed').
(c) All units in the target population have a positive probability of being observed in each of the data sets, and for at least one data set, these inclusion probabilities are homogeneous.
(d) The data sets can be linked perfectly.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

(e) The data sets do not contain units that do not belong to the target population ('erroneous captures'), nor do they contain duplicates.

These assumptions are rather strong. Research has shown that estimates of population size based on the capture–recapture method can be severely biased when some of these assumptions are violated (Brown *et al*., 2006; Van der Heijden *et al.*, 2012; Gerritse, 2016).

There is ongoing research into generalisations of the capture–recapture method and alternative methods that require less strong assumptions. Assumptions (a) and (c) are often relaxed by adding covariates to the model. Here, a problem may be that some covariates are not available in all data sources. Incomplete covariates may be handled by maximum likelihood under a Missing At Random assumption; see Van der Heijden *et al.* (2018) for a recent discussion with applications. Lawless (2014, Chapter 17) discussed adaptations of the capture–recapture method to open populations (assumption (b)). Extensions that can account for linkage errors (assumption (d)) were developed by Ding and Fienberg (1994, 1996) and Di Consiglio and Tuoto (2015). De Wolf *et al*. (2018) provide a synthesis and further generalisation of these extensions. These methods work under probabilistic record linkage, by correcting the observed counts for bias due to erroneous and missed links.

Assumption (e) is violated in the presence of *overcoverage* in one or more data sets. Di Cecco *et al.* (2018) have developed an extended capture–recapture method that can account for overcoverage as well as data sets that contain certain specific subpopulations only (so that not all units in the target population have a positive probability of being observed in each of the data sets, and assumption (c) is violated). This approach is based on an LC model, with erroneous captures indicated by a latent variable. A practical drawback of this method is that it requires at least four linked data sets. An alternative approach for handling simultaneous undercoverage and overcoverage, which is not based on the capture–recapture method, was proposed by Zhang (2015).

Overcoverage is a wider problem that also occurs outside the context of capture–recapture methods. For instance, a population register may suffer from overcoverage due to delayed deregistration of inactive units. In practice, overcoverage and duplicated records are often handled by clerical review or by applying deterministic rules (Di Cecco *et al.*, 2018). Assessing the amount of overcoverage and its effects on estimates may be difficult in some applications, in particular, when overcoverage is caused by false positive linkage errors (Bakker, 2011b). In the context of a traditional census, the overcoverage rate is usually estimated from a post-enumeration survey. In a multi-source context, the overcoverage rate may be assessed by linking administrative or survey data from auxiliary sources to the main data set (UN/ECE, 2014, pp. 75–77).

4.6 Aggregated Data Only

Situation 6 (see Figure 6) is the macro data counterpart of Situation 4: in Situation 6, only aggregated data overlap with each other and need to be reconciled. An example of Situation 6 is provided by the National Accounts, where aggregated data from different data sets need to be reconciled with each other subject to both equality and inequality constraints.

To reconcile aggregated data, macro-integration can be used (see, e.g. Mushkudiani *et al*. 2012). When macro-integration is applied, only estimated figures at an aggregated level are adjusted. The goals of macro-integration are to obtain a more accurate, numerically consistent and complete set of estimates for the variables of interest.

Often, the starting point of macro-integration is a set of estimates in tabular form. The entries of the tables are adjusted so all differences between tables are reconciled, and the entries with the highest variance are adjusted the most.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

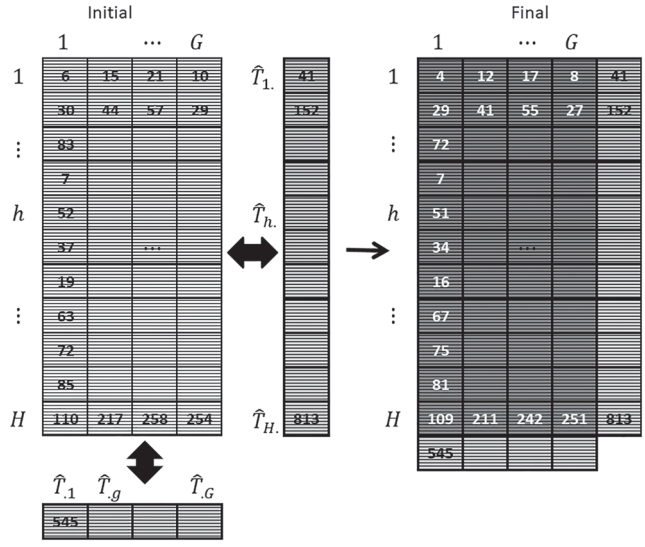

Figure 6. *Combining macro data sets.*

In the macro-integration approach, often a constrained optimisation problem is constructed. A target function, for instance, a quadratic form of differences between the original and the adjusted values, is minimised, subject to the constraints that the adjusted common figures in different tables are equal to each other and additivity of the adjusted tables is maintained. Inequality constraints can be imposed on these quadratic optimisation problems. In the literature, Bayesian macro-integration methods have also been proposed. Several methods for macro-integration have been developed, see, for instance, Stone *et al*. (1942), Byron (1978), Sefton and Weale (1995), Magnus *et al*. (2000), Boonstra *et al*. (2011), Mushkudiani *et al*. (2012; 2015) and Daalmans (2015).

Macro-integration can reconcile several tables simultaneously, as long as the number of variables or constraints does not become too large. With current software and computers, problems with several hundred thousand unknowns and constraints can nowadays be solved.

Macro-integration can only be applied for correcting random errors, not for correcting systematic errors as application to systematic errors is likely to lead to biased results. Systematic errors, especially large ones, have to be corrected by another approach, for example, by manual data editing, before macro-editing can be applied successfully.

When one wants to use macro-integration, it is important that (an approximation to) the variance of each entry in the tables to be reconciled is available, can be computed or can somehow be approximated. In some cases, one may have to rely on expert knowledge in order to approximate these variances (see, e.g. Xie *et al.*, 2018).

In practice, results after macro-integration of large sets of tables, such as National Accounts, are checked manually for plausibility, for instance, by inspecting time series of reconciled figures. If needed, the reconciliation is repeated after removing some errors overlooked in the first instance.

4.7 Micro Data and Aggregated Data

Situation 7 (see Figure 7) is characterised by a variation on Situation 4, by which aggregated data are available besides micro data. There is still overlap between the data sets, from which the need arises to reconcile the statistics at some aggregated level. Of particular interest here

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

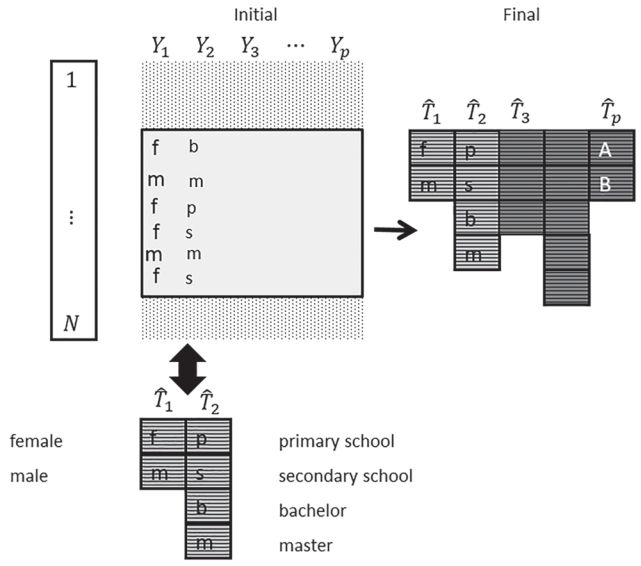

Figure 7. *Combining a micro data set with a macro data set.*

is the case that the aggregated data are estimates themselves. Otherwise, the reconciliation can be achieved by means of calibration which is a standard approach in survey sampling (see, e.g. Chapter 6 in Sarndal *et al*., 1992). In Figure 7, the aggregated data are denoted by $\hat{T}_1, \ldots, \hat{T}_p$ to highlight that in practice, these are often estimated population totals.

We assume that several tables have to be estimated using the available micro data and aggregated data. An example of Situation 7 is the Dutch Population census, which is based on a mix of administrative data sets and sample survey data as mentioned before. Population totals, either known from an administrative data set or previously estimated, are imposed as benchmarks provided they overlap with an additional survey data set that is needed to produce new output statistics.

When micro data and aggregated data have to be reconciled, several methods are available, such as repeated weighting, repeated imputation, mass imputation and macro-integration (see also De Waal, 2016). In repeated weighting, population tables are estimated sequentially. Data from a data set covering the entire population can simply be counted. Data only available from surveys are weighted. A separate set of weights is assigned to survey units for each table of population totals to be estimated. When estimating a new table, all cell values and margins of this table that are known or have already been estimated for previous tables are kept fixed. This is achieved by using regression weighting to calibrate to these known or previously estimated values (Houbiers, 2004). This ensures numerical consistency of the cell values and margins of the new table and previous estimates, if calibration weights can be found. That such calibration weights can be found is not guaranteed, however. Repeated weighting is mainly applied to ensure numerical consistency between estimated tables. However, calibrating to totals based on large sample sizes generally leads to a reduction of the sample variance for tables based on smaller sample sizes (see, e.g. Houbiers, 2004).

A strong aspect of repeated weighting is that (statistical and logical) relationships between data items from a single data source are automatically maintained. The occurrence of empty cells in high-dimensional tables, that is, cells without any observations, complicates the use of repeated weighting as weighting empty cells leads to population estimates with value zero. In

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

some cases, either very large or very small weights may then have to be given to other cells in order to preserve known or previously estimated values. In other cases, it may not even be possible to find suitable weights at all.

Repeated imputation is similar to repeated weighting. Repeated imputation is again a sequential approach where tables are estimated one by one. For some variables in a table, estimates may have already been produced while estimating a previous table. These variables are then calibrated to the previously estimated values by applying an imputation method that preserves known or previously estimated values. For each new table to be estimated, a new imputation model is constructed.

The occurrence of empty cells is usually not a serious problem for these imputation methods. However, with repeated imputation it may be difficult to preserve relationships between variables, even for variables occurring in the same data set. The results of both repeated weighting and repeated imputation depend on the order in which tables are estimated.

A prerequisite for applying repeated imputation is an imputation method that succeeds in preserving the statistical aspects of the true data as well as possible and that is able to preserve previously estimated values. Preferably, the imputation method should also satisfy edit restrictions on the data. Such imputation methods have been developed by, for instance, Chambers and Ren (2004), Zhang (2008), Zhang and Nordbotten (2008), Pannekoek *et al*. (2013), Coutinho *et al*. (2013), Kim *et al*. (2014), Da Silva and Zhang (2014) and De Waal *et al*. (2017a). Which imputation method is most appropriate depends on the kind of data (e.g. numerical versus categorical data), the missing data mechanism and the aims one tries to fulfil (e.g. should logical rules, such as that males cannot be pregnant, be fulfilled at the micro level?).

When mass imputation is used, one imputes all fields for which no value was observed for all population units. Mass imputation hence leads to a data set with values for all variables and all units. After imputation, estimates for population totals can be obtained by simply counting or summing the values of the corresponding variables.

The major risk of mass imputation is that the mass-imputed data may be used to estimate or analyse aspects that were not accounted for in the imputation model. The results of such an estimation or analysis procedure are likely to be biased. It is generally impossible to capture all relevant variables and relations in the imputation model, simply because there are not enough observations to estimate all model parameters accurately, which implies that many relations found in the imputed data will not reflect the relations in the population. Note that this is not necessarily a problem for repeated imputation. In that case, a separate imputation model, involving a limited number of variables only, is constructed for each new table. Mass imputation has, for instance, been studied by Whitridge *et al*. (1990), Whitridge and Kovar (1990) and Shlomo *et al*. (2009).

Macro-integration has already been described for Situation 6 and can be applied in Situation 7 too by first transforming the micro data to aggregated data themselves. As the transformation is usually carried out by means of weighting the data, empty cells may complicate the procedure, just like for repeated weighting. A (potential) drawback of the macro-integration approach in Situation 7 is that one cannot re-calculate the adjusted table figures from the underlying micro data directly. This problem may in some cases be overcome by deriving weights by means of the calibration estimator, using the reconciled macro-integrated figures to calibrate the results. Such weights do not necessarily exist, however.

An advantage of macro-integration over repeated weighting and repeated imputation is that all tables to be estimated can be produced simultaneously. So, whereas the results of repeated weighting and repeated imputation are order dependent, the results of macro-integration are not. Besides, the simultaneous estimation of all tables may lead to more accurate estimates. In summary, what is the most suitable method for reconciliation of micro data on macro data depends

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

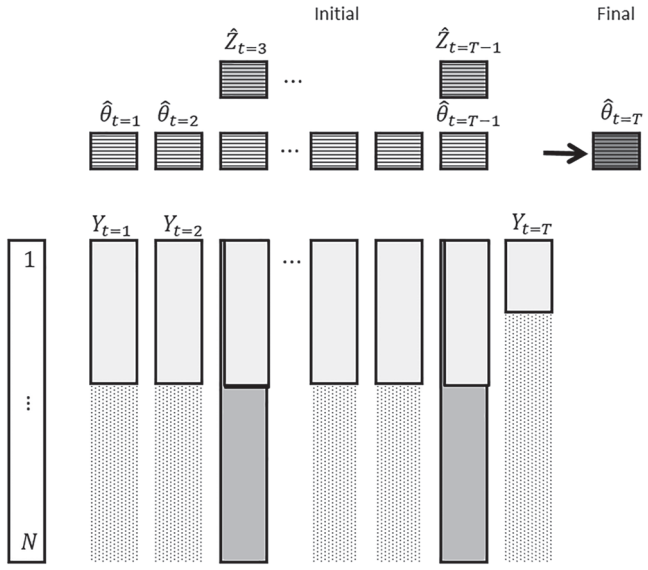

Figure 8. *Combining longitudinal data sets.*

on the properties of the data and on the targeted results. The answer depends on questions such as: is it important that the macro estimates can be directly (re-)calculated from the micro data, are there many empty cells, do logical relations play a role, and will the micro data be used by other researchers?

4.8 Longitudinal Data

Finally, *longitudinal data* are introduced in Situation 8. We limit ourselves to the issue of reconciling a time series of high frequency with one of a low frequency, as illustrated in Figure 8. The difference with the macro-integration in Situation 6 is that the data are now related to each other over time. The data of the low-frequency series are usually considered to be exogenous and are kept fixed, because these are usually based on the most comprehensive information.

When a high-frequency series is adjusted to have temporal consistency with a low-frequency series of the same variable, usually measured from a different data source, this is known as *benchmarking* (European Commission, 2018, p. 7). A related problem is that of *disaggregation*: a series of low frequency of a target variable is disaggregated by using an indicator series of high frequency for the target variable (European Commission, 2018, p. 7).

Situation 8 is for instance found at Statistics Netherlands where monthly turnover based on a sample survey of enterprises is used to compute turnover indices for the short-term statistics. These indices are computed for a number of publication cells. An example of the time series of the publication cell 'Manufacture of cutlery, tools and general hardware', from January 2010 till December 2011 is given in Figure 9. These sample survey data (labelled as 'source' in Figure 9) are benchmarked against quarterly turnover values. The horizontal lines in Figure 9 represent the average monthly index values per quarter of the source and the benchmark data. The quarterly benchmark turnover values are largely based on VAT data supplemented by survey data, which was explained already in the example for Situation 2. These quarterly data are kept fixed, because they cover nearly the complete population.

A wide range of methods is available for benchmarking. Perhaps the most basic method is to preserve the original levels with prorating. Prorating means that the level estimates are adjusted

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

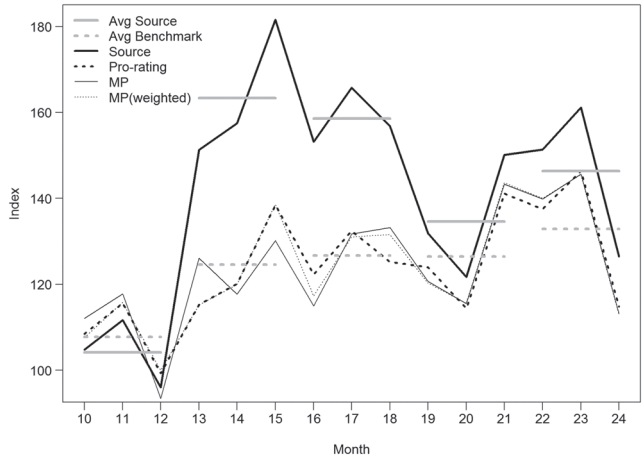

Figure 9. *Index of monthly turnover: source data and three benchmarked series: 'Prorating', 'MP' (movement preservation) and 'MP (weighted)' (see text). Month 1 = January 2010.*

with the same relative factor. Another method to preserve the original levels is that by Chow and Lin (1971). It expresses the estimation of the high-frequency values as a linear regression on the low-frequency values and finds the solution by generalised least squares.

A disadvantage of prorating and of the Chow–Lin method is that they lead to the so-called step problem: when observing reconciliation adjustments of the changes between two successive high-frequency periods, disproportionally large adjustments may be observed in the transition from one low-frequency period to the next. For instance, in the turnover example, the monthly growth rate in January 2011 was 57.5% in the source data, and after applying prorating, it was adjusted to 16.1% due to the step problem (Figure 9). A similarly large adjustment can be seen in the growth rate of July 2011.

An alternative to level preservation is movement preservation (MP). MP methods aim to preserve the changes in the original high-frequency series. Examples of methods in this class are the ones by Denton (1971), their slightly modified variants by Cholette (1984) and the extensions of Chow–Lin by Fernández (1981).

In order to give a more formal presentation of benchmarking, let $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ stand for the values of a monthly time series and let $\mathbf{b} = (b_1, b_2, \ldots, b_m)'$ be the values of a quarterly time series, which is kept fixed. Denote the benchmarked values by $\mathbf{x}^*$. After benchmarking, it should hold that $\sum_{j=3(q-1)+1}^{3q} x_j^* = b_q$. The *additive first-order Denton method* finds benchmarked values by minimising the squared differences between adjusted and original first-order differences over the entire period of the series (Bikker *et al*., 2011), more formally stated by

$$\min_{x_j^*} \sum_{j=1}^{n} \left(\Delta x_j^* - \Delta x_j\right)^2 \text{ with } \Delta x_j = x_j - x_{j-1} \text{ and } \Delta x_1 = x_1.$$

Therefore, the benchmarked values are determined not only by the corresponding quarters but also by previous and next quarters. This way, a large shift in monthly changes just before and after the end of a quarter is avoided. In the turnover example, the monthly growth rate in January 2011 for the series benchmarked by the MP approach was 35%, which is closer to the growth rate of the source than was the case after benchmarking with prorating. Also, the growth rate adjustment in July 2011 was smaller after applying the MP approach than after prorating.

Benchmarking can also be applied to multiple time-related variables. The problem now is to deal with time constraints and with cross-sectional constraints between variables (Bikker *et*

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

al., 2011). Di Fonzo and Marini (2003; 2005) and Bikker and Buijtenhek (2006) combined the Denton method for time constraints with the method of Stone *et al*. (1942) for handling cross-sectional constraints between the variables.

A multivariate benchmarking method can be refined by applying weights to the adjustments made to each series. These weights should reflect the relative accuracy of the estimated growth rates of the high level frequency series. Usually, growth rates of reliably measured series are preserved more strongly than the growth rates of inaccurately measured series.

Bikker *et al*. (2013) extended the method to include other modelling features, such as constraints that have to be satisfied only approximately (soft constraints), ratio constraints and inequality constraints.

The reconciliation methods in this section cannot be used for data with (large) systematic errors, because of a smearing effect: an error in one value contaminates other values' estimates. Hence, it is important to check the time series for large systematic errors and to correct those before applying benchmarking. This is usually carried out interactively by confronting the preliminary data with the constraints.

After benchmarking, one should always inspect the corrections to judge the plausibility of results. Guidelines on how to apply benchmarking in specific situations can be found in European Commission (2018).

5 Discussion

We are fully aware that the basic situations we have considered in this paper do not offer a complete description of all situations that may arise in practice and that our basic situations give a simplified view of reality. At the same time, we do feel that this paper offers useful guidelines to producers of multi-source statistics. Many situations arising in practice are variations of the basic situations that we have discussed in this paper or combinations of such basic situations. The basic situations and the corresponding methods we discussed in this paper should at least give producers of multi-source statistics a good starting point to handle such cases. For instance, when we are dealing with a combination of two basic situations, a logical starting point would be to consider using methods for these two situations in combination. As an example, for multi-source data with undercoverage and a common target variable with measurement for overlapping units, one could consider using capture–recapture techniques (Section 4.5) in combination with LC models (Section 4.4). This is indeed the approach taken at Statistics Netherlands.

In the discussion of the basic situations, we have pinpointed important issues that can occur for these situations. This will allow producers of multi-source statistics to anticipate the problems that may occur for their specific situation. In the discussion of the basic situations, we also described and gave references to important methods that can be used to overcome the problems. Hopefully, this will give the producers of multi-source statistics a flying start to overcome the problems for their own specific case. Many of the methods referred to in this paper have only recently been developed. These methods are therefore still in their infancy and will hopefully be improved upon in many different aspects in the coming years.

Finally, we remark that after combining data sets, one is usually interested in estimating the accuracy of the outcomes. Different quality measures and methods to compute them for various situations are currently under development for this purpose in the *ESSnet on Quality of Multi-source Statistics*, which is partly funded by the EU (see, e.g. De Waal *et al*., 2017b).

We hope that our discussion of various situations will inspire other researchers to do research on the highly important and interesting area of producing multi-source statistics.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Acknowledgements

The authors thank the referees and the co-editor-in-chief for their comments that led to considerable improvements of the article.

References

Baffour, B., Brown, J. J. & Smith, P. W. F. (2013). An investigation of triple system estimators in censuses. *Stat. J. Int. Assoc. Off. Stat.*, **29**, 53–68.

Bakker, B. F. M. (2011a). *Micro-Integration: State of the Art*. Chapter 5 in: *State of the Art on Statistical Methodologies for Data Integration*. Report on WP1 of the ESS net on Data Integration.

Bakker, B. F. M. (2011b). *Micro Integration*, Statistical Methods (201108).The Hague/Heerlen: Statistics Netherlands.

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. *Statist. Neerlandica*, **66**, 8–17.

Bakker, B. F. M. & Daas, P. (2012). Some methodological issues of register based research. *Statist. Neerlandica*, **66**, 2–7.

Biemer, P. P. (2011). *Latent Class Analysis of Survey Error.* Hoboken, New Jersey: John Wiley & Sons.

Bikker, R. P. & Buijtenhek, S. (2006). *Alignment of Quarterly Sector Accounts to Annual Data* Voorburg: Statistics Netherlands. http://www.cbs.nl/NR/rdonlyres/D918B487-45C7-4C3C-ACD0-oE1C86E6CAFA/0/Benchmarking_QSA.pdf.

Bikker, R., Daalmans, J. & Mushkudiani, N. (2011). *Macro-integration, Data Reconciliation*, Statistical Methods (201104). The Hague/Heerlen: Statistics Netherlands.

Bikker, R., Daalmans, J. & Mushkudiani, N. (2013). Benchmarking large accounting frameworks: a generalised multivariate model. *Econ. Syst. Res.*, **25**, 390–408.

Bishop, Y., Fienberg, S. & Holland, P. (1975). *Discrete Multivariate Analysis, Theory and Practice.* New York: McGraw-Hill.

Blackwell, M., Honaker, J. & King, G. (2017). A unified approach to measurement error and missing data: Overview and applications. *Sociol. Methods Res.*, **46**, 303–341.

Boeschoten, L., Oberski, D. & De Waal, T. (2017). Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *J. Off. Stat.*, **33**, 921–962.

Boeschoten, L., Oberski, D., De Waal, T. & Vermunt, J. K. (2018). Updating latent class imputations with external auxiliary variables. *Struct. Equ. Model.*, **25**, 750–761.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: John Wiley & Sons.

Boonstra, H. J., De Blois, C. J. & Linders, G. J. (2011). Macro-integration with inequality constraints: an application to the integration of transport and trade statistics. *Statist. Neerlandica*, **65**, 407–431.

Brown, J. J., Abott, O. & Diamond, I. D. (2006). Dependence in the 2001 one-number census project. *J. R. Stat. Soc. A. Stat. Soc.*, **169**, 883–902.

Brown, J., Diamond, I., Chambers, R., Buckner, L. & Teague, A. (1999). A methodological strategy for a one-number census in the UK. *J. R. Stat. Soc. A. Stat. Soc.*, **162**, 247–267.

Byron, R. P. (1978). The estimation of large social account matrices. *J. R. Stat. Soc. A*, **141**, 359–367.

Chambers, R. L. & Ren, R. (2004). Outlier robust imputation of survey data. In *ASA Proceedings of the Joint Statistical Meetings*, pp. 3336–3344. Toronto: American Statistical Association.

Chen, C., Page, M. J. & Stewart, J. M. (2016). Creating new and improved business statistics by maximising the use of administrative data. In *Fifth International Conference on Established Surveys*, Geneva, Switzerland.

Cholette, P. (1984). Adjusting sub-annual series to yearly benchmarks. *Surv. Methodol.*, **10**, 35–49.

Chow, G. C. & Lin, A. (1971). Best linear unbiased interpolation, and extrapolation of time series by related series. *Rev. Econ. Stat.*, **53**, 372–375.

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. In *Data-Centric Systems and Applications*. Berlin Heidelberg: Springer-Verlag.

Conti, P. L., Marella, D. & Neri, A. (2017). Statistical matching and uncertainty analysis in combining household income and expenditure data. *Stat Methods Appl*, **26**, 485–505.

Coutinho, W., De Waal, T. & Shlomo, N. (2013). Calibrated hot deck imputation subject to edit restrictions. *J. Off. Stat.*, **29**, 299–321.

D'Orazio, M., Di Zio, M. & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* Chichester, UK: John Wiley and Sons.

Da Silva, D. N. & Zhang, L. -C. (2014). Adjustments for survey imputed datasets to achieve first and second-order properties. In *ASA Proceedings of the Joint Statistical Meetings*, Boston.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Daalmans, J. (2015). *Estimating Detailed Frequency Tables from Registers and Sample Surveys*, Discussion paper The Hague: Statistics Netherlands.
Daas, P. J. H., Puts, M. J., Buelens, B. & Van den Hurk, P. A. M. (2015). Big data as a source for official statistics. *J. Off. Stat.*, **31**, 249–262.
De Waal, T. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Stat. J. IAOS*, **32**, 231–243.
De Waal, T., Coutinho, W. & Shlomo, N. (2017a). Calibrated hot deck imputation for numerical data under edit restrictions. *J. Surv. Stat. Methodol.*, **5**, 372–397.
De Waal, T., Pannekoek, J. & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation.* New York: John Wiley & Sons.
De Waal, T., Van Delden, A. & Scholtus, S. (2017b). Output quality of multi-source statistics. In *Paper Presented at the NTTS Conference*, Brussels.
De Wolf, P.-P., Van der Laan, J. & Zult, D. (2018). *Connecting Correction Methods for Linkage Error in Capture–Recapture*, Discussion Paper. The Hague: Statistics Netherlands.
Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization. *J. Am. Stat. Assoc.*, **66**, 99–102.
Di Cecco, D., Di Zio, M., Filipponi, D. & Rocchetti, I. (2018). Population size estimation using multiple incomplete lists with overcoverage. *J. Off. Stat.*, **34**, 557–572.
Di Consiglio, L. & Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *J. Off. Stat.*, **31**, 415–429.
Di Fonzo, T. & Marini, M. (2003). *Benchmarking Systems of Seasonally Adjusted Time Series According to Denton's Moving Preservation Principle*: University of Padua. http://www.oecd.org/dataoecd/59/19/21778574.pdf.
Di Fonzo, T. & Marini, M. (2005). *Benchmarking a System of Time Series: Denton's Movement Preservation Principle vs. Data Based Procedure*: University of Padova. http://epp.eurostat.cec.eu.int/cache/ITY_PUBLIC/KSDT-05-008/EN/KS-DT-05-008-EN.pdf.
Di Zio, M. & Luzi, O. (2014). Theme: Editing Administrative Data. In *Memobust Handbook on Methodology for Modern Business Statistics*. Luxembourg: Eurostat.
Ding, Y. & Fienberg, S. E. (1994). Dual system estimation of census undercount in the presence of matching error. *Surv. Methodol.*, **20**, 149–158.
Ding, Y. & Fienberg, S. E. (1996). Multiple sample estimation of population and census undercount in the presence of matching errors. *Surv. Methodol.*, **22**, 55–64.
Enderer, J. (2008). Is the utilization of administrative data in short term statistics an ideal standard in the conflicting priorities of user demands, response burden and budget restrictions? In *Proceedings of the IAOS Conference 'Reshaping Official Statistics'*, Shanghai.
European Commission. (2018). *ESS Guidelines on Temporal Disaggregation, Benchmarking and Reconciliation. From Annual to Quarterly to Monthly Data*. Report from the Task Force Temporal Disaggregation. Available at https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-06-18-355?inheritRedirect=true&redirect=%2Feurostat%2Fpublications%2Fmanuals-and-guidelines.
Eurostat. (2015). *ESS Handbook for Quality Reports. Eurostat Manuals and Guidelines* Luxembourg: Eurostat.
Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *J. Am. Stat. Assoc.*, **64**, 1183–1210.
Fernández, R. B. (1981). A methodological note on the estimation of time series. *Rev. Econ. Stat.*, **63**, 471–476.
Fienberg, S. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 409–439.
Gerritse, S. C. (2016). *An Application of Population Size Estimation to Official Statistics*, PhD Thesis, Utrecht University.
Gerritse, S. C., Bakker, B. F. M., De Wolf, P. P. & Van der Heijden, P. G. M. (2016). *Undercoverage of the Population Register in the Netherlands 2010*. Published as Chapter 5 in Gerritse (2016).
Guarnera, U. & Varriale, R. (2015). Estimation and editing for data from different sources. An approach based on latent class model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing. Budapest.
Guarnera, U. & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Stat. J. IAOS*, **32**, 537–544.
Hagenaars, J. A. & McCutcheon, A. L. (eds). (2002). *Applied Latent Class Analysis.* New York: Cambridge University Press.
Harron, K., Goldstein, H. & Dibben, C. (2016). *Methodological Developments in Data Linkage.* Chichester: Wiley.
Herzog, T. N., Scheuren, F. J. & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques.* New York: Springer.
Houbiers, M. (2004). Towards a social statistical database and unified estimates at Statistics Netherlands. *J. Off. Stat.*, **20**, 55–75.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

International Working Group for Disease Monitoring and Forecasting. (1995). Capture–recapture and multiple record systems estimation. Part 1. History and theoretical development. *Am. J. Epidemiol.*, **142**, 1059–1068.
Kim, J. K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation. *Surv. Methodol.*, **42**, 19–40.
Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H. & Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *J. Bus. Econ. Stat.*, **32**, 375–386.
Landefeld, S. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues. In *International Conference on Big Data for Official Statistics*, Beijing, China.
Lawless, F. (2014). *Statistics in Action. A Canadian Outlook.* Ontario: Apple Academic Press Inc.
Linder, F., Van Roon, D. & Bakker, B. F. M. (2011). Combining data from administrative sources and sample surveys: the single variable case. In *ESSnet Data Integration, WP4 Case Studies*. Luxembourg: Eurostat, pp. 39–97.
Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.
Lohr, S. L. & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Stat. Sci.*, **32**, 293–312.
Magnus, J. T., Van Tongeren, J. W. & De Vos, A. F. (2000). National accounts estimation using indicator ratios. *Rev. Income Wealth*, **46**, 329–350.
McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models.* New York: John Wiley & Sons.
Meijer, E., Rohwedder, S. & Wansbeek, T. (2012). Measurement error in earnings data: using a mixture model approach to combine survey and register data. *J. Bus. Econ. Stat.*, **30**, 191–201.
Mushkudiani, N., Daalmans, J. & Pannekoek, J. (2012). *Macro-integration Techniques with Applications to Census Tables and Labour Market Statistics*, Discussion paper. Statistics Netherlands: The Hague.
Mushkudiani, N., Daalmans, J. & Pannekoek, J. (2015). Reconciliation of labour market statistics using macro-integration. *Stat. J. IAOS*, **31**, 257–262.
Nordbotten, S. (2010). The use of administrative data in official statistics—past, present, and future—with special reference to the nordic countries. In *Official Statistics, Methodology and Applications in Honour of Daniel Thorburn*, Eds. Carlson, Nyquist & Villani. Stockholm, Sweden: Stockholm University, pp. 205–223. Available at officialstatistics.wordpress.com.
Oberski, D. (2017). Estimating error rates in an administrative register and survey questions using a latent class model. In *Total Survey Error in Practice*, pp. 341–358. New York: Wiley.
Ouwehand, P. & Schouten, B. (2014). Measuring representativeness of short-term statistics. *J. Off. Stat.*, **30**, 623–649.
Pannekoek, J., Shlomo, N. & De Waal, T. (2013). Calibrated imputation of numerical data under linear edit restrictions. *Ann. Appl. Stat.*, **7**, 1983–2006.
Pavlopoulos, D. & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Surv. Methodol.*, **41**, 197–214.
Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.
Särndal, C. E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.
Scholtus, S., Bakker, B.F.M. & Van Delden, A. (2015). *Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables*, Discussion paper. The Hague: Statistics Netherlands.
Sefton, J. & Weale, M. (1995). *Reconciliation of National Income and Expenditure.* Cambridge, UK: Cambridge University Press.
Shlomo, N., De Waal, T. & Pannekoek, J. (2009). *Mass Imputation for Building a Numerical Statistical Database.* Neuchâtel, Switzerland: UN/ECE Work Session on Statistical Data Editing.
Si, Y. & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.*, **38**, 499–521.
Singh, A. C., Mantel, H. J., Kinack, M. D. & Rowe, G. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Surv. Methodol.*, **19**, 59–79.
Stone, R., Champernowne, D. G. & Meade, J. E. (1942). The precision of national income estimates. *Rev. Econ. Stud.*, **9**, 111–125.
Stuart, A. & Ord, J. K. (1991). *Kendall's Advanced Theory of Statistics, Volume 2*, Fifth. London: Edward Arnold.
UN/ECE. (2014). *Measuring Population and Housing. Practices of UNECE Countries in the 2010 Round of Censuses* New York and Geneva: United Nations.
Van Delden, A. & De Wolf, P. P. (2013). A production system for quarterly turnover levels and growth rates based on VAT data. In *Proceedings of the Conferences on New Techniques and Technologies for Statistics*, Brussels. Available at http://www.cros-portal.eu/sites/default/files//NTTS2013%20Proceedings_0.pdf (accessed December 2013).
Van Delden, A., Du Chatinier, B. & Scholtus, S. (2019). Accuracy in the Application of Statistical Matching Methods for Continuous Variables using Auxiliary Data. *J. Surv. Stat. Methodol.* Accepted for publication.
Van Delden, A., Lorenc, B., Struijs, P. & Zhang, L. -C. (2018a). On statistical unit errors in business statistics. *Lett. Ed. J. Off. Stat.*, **34**, 573–580.
Van Delden, A., Pannekoek, J., Banning, R. & De Boer, A. (2016). Analysing correspondence between administrative and survey data. *Stat. J. IAOS*, **32**, 569–584.

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.

Van Delden, A., Van der Laan, J. & Prins, M. J. (2018b). Detecting reporting errors in data from decentralised, autonomous, administrations with an application to hospital data. *J. Off. Stat.*, **34**, 863–888.
Van der Heijden, P. G. M., Smith, P. A., Cruyff, M. & Bakker, B. (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *J. Off. Stat.*, **34**, 239–263.
Van der Heijden, P. G. M., Whittaker, J., Cruyff, M., Bakker, B. & Van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: invariant population size estimates and the role of active and passive covariates. *Ann. Appl. Stat.*, **6**, 831–852.
Whitridge, P., Bureau, M. & Kovar, J. (1990). Mass Imputation at Statistics Canada. In *Proceedings of the Annual Research Conference*: U.S. Census Bureau, Washington D.C., pp. 666–675.
Whitridge, P. & Kovar, J. (1990). Use of mass imputation to estimate for subsample variables. In *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, Anaheim, California, pp. 132–137.
Wolter, K. M. (1986). Some coverage error models for census data. *J. Am. Stat. Assoc.*, **81**, 338–346.
Xie, Y., Lennmalm, A., Lennartsson, D. & De Groote, A. (2018). Uncertainty and automatic balancing of national accounts with a Swedish application. *J. IAOS*, **34**, 263–269.
Zhang, L. C. (2008). *A Triple-Goal Imputation Method for Statistical Registers* Neuchâtel: UN/ECE Work Session on Statistical Data Editing.
Zhang, L. C. (2011). A unit-error theory for register-based household statistics. *J. Off. Stat.*, **27**, 415–432.
Zhang, L. C. (2012). Topics of statistical theory for register-based statistics and data integration. *Stat. Neerlandica*, **66**, 41–63.
Zhang, L. C. (2014). Data integration. *Surv. Stat.*, **70**, 15–24.
Zhang, L. C. (2015). On modelling register coverage errors. *J. Off. Stat.*, **31**, 381–396.
Zhang, L. C. & Nordbotten, S. (2008). *Prediction and Imputation in ISEE: Tools for More Efficient Use of Combined Data Sources.* Vienna: UN/ECE Work Session on Statistical Data Editing.

[*Received June 2017, accepted July 2019*]

© 2019 The Authors. International Statistical Review © 2019 International Statistical Institute.